Structure and variability of recently inserted Alu family members

Mark A.Batzer[1,*], Gail E.Kilroy[1], Pamela E.Richard[1], Tamim H.Shaikh[1], Trent D.Desselle[1], Carol L.Hoppens[1] and Prescott L.Deininger[1,2]

[1]Department of Biochemistry and Molecular Biology, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112 and [2]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, LA 70121, USA

Received September 28, 1990; Revised and Accepted November 1, 1990

EMBL accession nos X54175, X55922–X55933

ABSTRACT

The HS subfamily of Alu sequences is comprised of a group of nearly identical members. Individual subfamily members share 97.7% nucleotide identity with each other and 98.9% nucleotide identity with the HS consensus sequence. Individual subfamily members are on the average 2.8 million years old, and were probably derived from a single source 'master' gene sometime after the human/great ape divergence. The recent Alu family member insertions provide a better image of the structure of Alu retroposons before they have had the opportunity to change significantly. All of the HS subfamily members are flanked by perfect direct repeats as a result of insertion at staggered nicks. The 'master' gene from which the HS subfamily members were derived had an oligo-dA rich tail at least 40 bases long. The 'master' gene is very rich in CpG dinucleotides, but nucleotide substitutions within subfamily members accumulated in a random manner typical for Alu sequences with CpG substitutions occurring 9.2 fold faster than non-CpG substitutions.

INTRODUCTION

The Alu family of short interspersed repetitive DNA elements (SINEs) is only found within the genome of primates (for reviews see 1–3), having arisen within the last sixty-five million years (4). The Alu family represents one of the most successful classes of mobile elements, having amplified to a copy number in excess of 500,000 within the human genome (5). Alu sequences are distributed on average about 5000 bp apart within the human genome, but have also been found to cluster within specific genomic loci (6,7). Each Alu element is about 300 bp in length consisting of two tandemly arranged halves, with the right half containing an additional 31 bp relative to the left half (5). Alu sequences are ancestrally derived from the 7SL RNA gene (8). Alu elements contain a middle A-rich region, 3′ oligo-dA tail which is variable in length, and are flanked by short direct repeats which form during integration. Mobilization of Alu elements is thought to occur via an RNA polymerase III derived transcript in a process termed retroposition (9).

The Alu sequences distributed throughout primate genomes may be subdivided into groups of related subfamily members based on nucleotide divergence (10). Several laboratories have divided Alu sequences into different subfamilies which appear to have arisen within primate genomes at different times (11–15). The most recently formed subfamily of Alu sequences found within the human genome was originally referred to as the 'new' subfamily (16). It has subsequently been further characterized and termed the Predicted Variant (PV) subfamily (17) and the Human Specific (HS) subfamily (18). We will utilize the HS nomenclature. Interestingly, the HS subfamily (PV) was found to be transcritionally active, *in vivo* (17), a property necessary for an active retroposon. There are an estimated 500 (18) to 2000 (17) subfamily members in the human genome. Individual HS subfamily members share 5 diagnostic nucleotide substitutions (compared to older Alu sequences) as well as a high degree of nucleotide identity with the consensus sequence suggesting that they were derived from a single, or at most a closely related set of, source gene(s) (16–18). Previously, several members of the HS subfamily were found to be present only within the human genome, and absent at orthologous positions in the genomes of other primates, indicating that most, if not all of the HS subfamily members had amplified within the human genome within the last 4–6 million years (18). In this report, we present a detailed structural analysis of a number of these HS subfamily members.

MATERIALS AND METHODS

Library construction, screening and DNA sequencing

A randomly sheared genomic library (5 kb inserts) was prepared from HeLa DNA in bacteriophage λ2 Zap II (Stratagene) (18). The library was screened with an HS Alu-specific probe (5′-CACCGTTTTAGCCGGGATGG-3′) at high stringency (65°C 6×SSC/0.05% sodium pyrophosphate) (18). Hybridizing clones were plaque purified, and excision subcloned using *Escherichia coli* XL1-Blue and M13 R408 as recommended by

[*] To whom correspondence should be addressed

```
                    10        20        30        40        50        60        70        80        90       100
                     .         .         .         .         .         .         .         .        *.      *  .
HS CON      GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGG
HS P1N5     ........C......................................T....................................................
HS P1N6     .......T...................G.............................A.....T....................................
HS C2N4     .........................................................A..........................................
HS C3N1     xxxxxxx.............................................................................................
HS C3N2     ...........................G........................................................................
HS C3N3     ....................................................................................................
HS C3N4     ........A...........................................................................................
HS C3N6     xxxxxxxxxT.........T....T...............T...........................................................
HS C3N7     ....................................................................................................
HS C4N2     .........................................................A..........................................
HS C4N4     .........................................................A..........................................
HS C4N6     ....................................................................................................
HS C4N8     .................................................................................................T..
HS G15N2    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx..................................................................
HS G18N1    ..........A............................................A............................................
HS G18N2    ..................................................................................G.................
HS G19N1    ....C...............................................................................................
HS H3N1     xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx..............................................................
MLVI-2      T...................................................................................................
TPA 25      ....................................................................................................
HS C4N5     ...T................................................................................................

                   110       120       130       140       150       160       170       180       190       200
                     .         .         .         .    *    .         .         .   *     .         .         .
HS CON      TGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGT
HS P1N5     .........................................................................A........................T.
HS P1N6     ............................................G....T..................................................
HS C2N4     ..........................................A......T..................................................
HS C3N1     ....................................................................................................
HS C3N2     ....................................................................................................
HS C3N3     ....................................................................................................
HS C3N4     .........................T..........................................................................
HS C3N6     ....................................................................................................
HS C3N7     ....................................................................................................
HS C4N2     .............................................................C......................................
HS C4N4     ..........................................A......T..................................................
HS C4N6     .................................................................................................T..
HS C4N8     ..........................................A.........................................................
HS G15N2    ....................................................................................................
HS G18N1    ....................................................................................................
HS G18N2    .................................................................................................T..
HS G19N1    ....................................................................................................
HS H3N1     ....................................................................................................
MLVI-2      ........x...............^...............x.........x.....x...........................................
TPA 25      ......................C.........x................................TG...............................A.
HS C4N5     ..........A...........C.........x................................T..................................

                   210       220       230       240       250       260       270       280       290
                     .         .         .      *  .         .         .         .         .         .
HS CON      GAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAA
HS P1N5     ......A.....T........T.....................................................T..............
HS P1N6     .....T......................................................A.............................
HS C2N4     ..........................................................................................
HS C3N1     ..........................................................................................
HS C3N2     T.............................A...........................................................
HS C3N3     ..........................................................................................
HS C3N4     ..........................................................................................
HS C3N6     ...................................................................A......................
HS C3N7     .............A.....x...............TG.......................A...xx..........A.............
HS C4N2     ............................................................................A.............
HS C4N4     ..........................................................................................
HS C4N6     .....................................T....................................................
HS C4N8     ..........................................G................T......................C.......
HS G15N2    ............T.............................................................................
HS G18N1    ..........................................................................................
HS G18N2    ...................................................................T......................
HS G19N1    ..........................................................................................
HS H3N1     ..........................................................................................
MLVI-2      ................xx........................................................................
TPA 25      ............................................................A.............................
HS C4N5     ............................................................................A.............
```

Figure 1. Nucleotide sequences flanking HS subfamily members. Nucleotide sequences flanking TPA 25 (19), MLVI-2 (20), HS C2N4, C3N1, C4N4, C4N5, C4N6, and C4N8 (18), as well as several other sequences reported here are shown. Nucleotides encompassed in the direct repeats are underlined. The length of the oligo-dA rich tail is denoted by an (A) and a subscript indicating the number of adenine residues.

Stratagene. Individual subfamily members were sequenced by standard dideoxy procedures on excised pBluescript SK(−) dsDNA using Sequenase (U.S. Biochemicals) and [α-^{35}S]-dATP from internal HS Alu-specific and flanking primers in both directions as previously described (18), according to the suppliers conditions. DNA sequences were aligned and computer analyzed using PC/GENE (Intelligenetics). The TPA 25 (19), Mlvi-2 (20)' and HS C2N4, C3N1, C4N4, C4N5, C4N6 and C4N8 (18) were previously reported. Subfamily members HS C4N2, P1N5, P1N6, C3N2, C3N3, C3N4, C3N6, C3N7, G15N2, G18N1, G18N2, G19N1, and H3N1 were assigned EMBL accession numbers X54175 and X55922–X55933 respectively.

RESULTS

Nucleotide identity of HS subfamily members

The alignment of 22 individual HS subfamily members is depicted in Fig. 1. Of these, the sequence of eight were previously reported (18). Nucleotide substitutions were divided into total, CpG, and non-CpG changes for further analysis and pairwise comparisons (Table I). Inspection of the individual HS subfamily members shows that they share a high degree of nucleotide identity. Total pairwise divergence values ranged from 0–5.5% (16/290 differences) with an average of 2.3%. Divergence from the HS subfamily consensus ranged from 0–2.7% (8/290 substitutions) with an average of 1.1% (Table I).

Comparison to the HS consensus sequence (Table II) shows that the rate of CpG substitutions to either TpG or CpA varied from 0–12.5% for individual Alu family members (3.9% average), while non-CpG positions ranged from 0–1.7% (0.4% average). Thus, the substitution rate at CpG positions was 9.2-fold higher than the rate at non-CpG positions relative to the HS consensus; in the discussion we consider the implications of this observation for germ line methylation of Alu sequences.

A total of 65 nucleotide substitutions (excluding deletions, insertions and 5′-truncations) occurred in the 21 HS subfamily members analyzed here (Fig. 1). Transitions accounted for 72% (47/65) of the observed substitutions while transversions accounted for the remaining 28% (18/65). The distribution of CpG and non-CpG substitutions did not significantly differ from a binomial or poisson distribution respectively (data not shown). Therefore the single base substitutions located throughout individual HS subfamily members appear to have occurred in a random manner aside from the bias for CpG positions.

Two regions (nt 65–108 and 167–197, Fig. 1) within the HS subfamily members showed relatively low levels of substitutions. Although the first region encompasses the B-box of the internal RNA polymerase III promoter, which might suggest a reason for selection of that region (21, 22), that is also the region around the oligonucleotide that was used to select these clones. Thus, this region would be expected to be biased for low divergence during the cloning. Studies involving folding of the Alu-like region of the 7SL RNA molecule have suggested that the region from 245–260 is involved in a unique secondary structure that base pairs with region 69–88 (23–25). Thus, this region too may be subject to selection, although it is not clear that the sampling here provides a significant level of resolution.

Age of HS subfamily members

The approximate age of individual HS subfamily members was determined using the number of informative (non-CpG) substitutions relative to the HS consensus, and a rate of evolution for primate intervening nucleotide sequences of 0.15% per million years (26) (Table II). CpG positions are considered uninformative and must be eliminated from this analysis due to a much faster 'clock' (above and (15, 27)). Previous studies to calibrate the molecular clock involved pseudogenes and intergenic regions which are generally depleted of CpG dinucleotides (15, 27 and 26 respectively). Thus it is most appropriate to base the age of the Alu family members on the non-CpG positions. Using this approximation, the predictedage of individual subfamily members varied from less than 2.7 million years old (HS C2N4, C3N1, C3N3, C4N4, C4N6, G15N2, H3N1) to 11.3 million years old (HS P1N5). This analysis shows that HS subfamily members inserted into the genome approximately 2.8 million years ago on average. The average time of insertion represents a much more reliable estimate than the insertion time of any single subfamily member, because of the small number of changes in the individual sequences.

Table I. Pairwise Comparisons of Alu HS Subfamily Members[1,2,3,4]

Subfamily Member	P1N5	P1N6	C2N4	C3N1	C3N2	C3N3	C3N4	C3N6	C3N7	C4N2	C4N4	C4N6	C4N8	G15N2	G18N1	G18N2	G19N1	H3N1	MLVI-2	TPA 25	C4N5
P1N6	16(10)																				
C2N4	11(7)	7(5)																			
C3N1	8(4)	8(6)	3(3)																		
C3N2	11(5)	9(7)	6(4)	3(1)																	
C3N3	8(4)	8(6)	3(3)	0(0)	3(1)																
C3N4	10(5)	10(7)	5(4)	2(1)	5(2)	2(1)															
C3N6	13(6)	13(8)	8(5)	5(2)	8(3)	5(2)	7(3)														
C3N7	15(8)	13(8)	10(7)	7(4)	10(5)	7(4)	9(5)	12(6)													
C4N2	11(6)	9(6)	4(3)	3(2)	6(3)	3(2)	5(3)	8(4)	8(4)												
C4N4	11(7)	7(5)	0(0)	3(3)	6(4)	3(3)	5(4)	8(5)	10(7)	4(3)											
C4N6	10(6)	10(8)	5(5)	2(2)	5(3)	2(2)	4(3)	7(4)	9(6)	5(4)	5(5)										
C4N8	13(7)	13(9)	6(4)	5(3)	8(4)	5(3)	7(4)	10(5)	12(7)	8(5)	6(4)	7(5)									
G15N2	7(3)	9(7)	4(4)	1(1)	4(2)	1(1)	3(2)	6(3)	8(5)	4(3)	4(4)	3(3)	6(4)								
G18N1	10(5)	10(7)	5(4)	2(1)	5(2)	2(1)	4(2)	7(3)	9(5)	5(3)	5(4)	4(3)	7(4)	3(2)							
G18N2	11(6)	11(8)	6(5)	3(2)	6(3)	3(2)	5(3)	8(4)	10(6)	6(4)	6(5)	3(2)	8(5)	4(3)	5(3)						
G19N1	9(4)	9(6)	4(3)	1(0)	4(1)	1(0)	3(1)	6(2)	8(4)	4(2)	4(3)	3(2)	6(3)	2(1)	3(1)	4(2)					
H3N1	8(4)	8(6)	3(3)	0(0)	3(1)	0(0)	2(1)	5(2)	7(4)	3(2)	3(3)	2(2)	5(3)	1(1)	2(1)	3(2)	1(0)				
MLVI-2	14(4)	14(6)	9(3)	6(0)	9(1)	6(0)	8(1)	11(2)	13(4)	9(2)	9(3)	8(2)	11(3)	7(1)	8(1)	9(2)	7(0)	6(0)			
TPA 25	14(6)	12(6)	9(5)	6(2)	9(3)	6(2)	8(3)	11(4)	11(4)	9(4)	9(5)	8(4)	11(5)	7(3)	8(3)	9(4)	7(2)	6(2)	12(2)		
C4N5	14(6)	14(8)	9(5)	6(2)	9(3)	6(2)	8(3)	11(4)	11(4)	7(2)	9(5)	8(4)	11(5)	7(3)	8(3)	9(4)	7(2)	6(2)	12(2)	6(4)	
HS CON	8(4)	8(6)	3(3)	0(0)	3(1)	0(0)	2(1)	5(2)	7(4)	3(2)	3(3)	2(2)	5(3)	1(1)	2(1)	3(2)	1(0)	0(0)	7(0)	3(2)	3(2)

[1] Total changes are followed by CpG changes in parenthesis.

[2] Only CpG changes to TpG or CpA were counted.

[3] 5′ truncations were not included.

[4] TPA 25 and C4N5 are compared to the HS-2 consensus (18).

Table II. Nucleotide divergence and Age of Alu HS subfamily members

Subfamily Member	CpG	substitutions[1]	non-CpG	substitutions[2]	Age (million years)[3]
HS P1N5	4	(8.3)	4	(1.7)	11.3
HS P1N6	6	(12.5)	2	(0.8)	5.3
HS C2N4	3	(6.3)	0	(0)	< 2.7
HS C3N1	0	(0)	0	(0)	< 2.7
HS C3N2	1	(2.1)	2	(0.8)	5.3
HS C3N3	0	(0)	0	(0)	< 2.7
HS C3N4	1	(2.1)	1	(0.4)	2.7
HS C3N6	2	(4.2)	3	(1.2)	8.0
HS C3N7	4	(8.3)	1	(0.4)	2.7
HS C4N2	2	(4.2)	1	(0.4)	2.7
HS C4N4	3	(6.3)	0	(0)	< 2.7
HS C4N6	2	(4.2)	0	(0)	< 2.7
HS C4N8	3	(6.3)	2	(0.8)	5.3
HS G15N2	1	(2.1)	0	(0)	< 2.7
HS G18N1	1	(2.1)	1	(0.4)	2.7
HS G18N2	2	(4.2)	1	(0.4)	2.7
HS G19N1	0	(0)	1	(0.4)	2.7
HS H3N1	0	(0)	0	(0)	< 2.7
MLVI-2	0	(0)	1	(0.4)	2.7
TPA 25[4]	2	(4.2)	1	(0.4)	2.7
HS C4N5[4]	2	(4.2)	1	(0.4)	2.7
Average	1.9 ± 1.6	(3.9 ± 3.3)	1.0 ± 1.1	(0.4 ± 0.4)	(2.8 ± 2.9)

[1] Includes only CpG positions which mutated to TpG or CpA, total changes followed by percent change.

[2] Does not include 5′ truncations and internal deletions or insertions, total changes followed by percent change.

[3] Using a rate of nucleotide substitutions of 0.15% per million years (26).

[4] Compared to the HS-2 subfamily consensus sequence (18).

HS P1N5	AAAAGGAATATGC	[Alu]	A_{17}	AAAAGGAATATGC
HS P1N6	AAGAAAATGTAG	[Alu]	A_{17}	G $(A)_5$G $(A)_9$G AAGAAAATGTA
HS C2N4	AAGAGCAGGTGGTT	[Alu]	A_{32}	AAGAGCAGGTGGTT
HS C3N1	AAAACTGCCCTGCTC	[Alu]	A_{17}	AAAACTGCCCTGCTC
HS C3N2	AAAATTCCTAAATCG	[Alu]	A_{20}	AAAATTCCTAAATCG
HS C3N3	AAAAAAATCCTCCTT	[Alu]	A_{17}	AAAAAAATCCTCCTT
HS C3N4	AAGACATGCCCAGGAG	[Alu]	A_{33}	AAGACATGCCCAGGAG
HS C3N6	CAGCACA	[Alu]	A_{35}	CTTGTTTGTGGTCAGCACA
HS C3N7	TGGATCC	[Alu]	A_{14}	GAAATTAAAGTCAACTGGAT
HS C4N2	AAGATTAGAATCTA	[Alu]	A_{37}	AAGATTAGAATCTA
HS C4N4	AGAAGATAATAACT	[Alu]	A_{25}	CAGAAGATAATAACT
HS C4N6	AGAATGCCATGTGAG	[Alu]	A_{25}	AGAATGCCATGTGA
HS C4N8	AAAAAATTGTTGG	[Alu]	AC $(A)_{19}$	AAAAAATTGTT
HS G15N2	AAGAAAATATTTC	[Alu]	A_{24}	AAGAAAATATTTC
HS G18N1	AGAAGTTAGGATAATG	[Alu]	A_{11}	AGAAGTTAGGATAATG
HS G18N2	AGAATTAGTCTGGAGA	[Alu]	A_{21}	AGAATTAGTCTGGAGA
HS G19N1	GAATTCAC	[Alu]	A_{14}	GAATTC
HS H3N1	AAAAAGTATACTTTC	[Alu]	A_{21}	AAAAAGTATACTTTC
MLVI-2	GAAAATGT	[Alu]	A_{26}	GAAAATGT
TPA 25	AGAACACTAC	[Alu]	A_{21}	AGAACACTAC
HS C4N5	AAGGAATCTCA	[Alu]	A_{25}	AAGGAATCTC

Figure 2. Nucleotide sequences flanking HS subfamily members. Nucleotide sequences flanking TPA 25 (19), MLVI-2 (20), HS C2N4, C3N1, C4N4, C4N5, C4N6, and C4N8 (18), as well as several other sequences reported here are shown. Nucleotides encompassed in the direct repeats are underlined. The length of the oligo-dA rich tail is denoted by an (A) and a subscript indicating the number of adenine residues.

HS subfamily member structure

Inspection of flanking nucleotide sequences shows that each HS subfamily member was flanked by short direct repeats (Fig. 2). The length of each set of direct repeats varied from 5 bp (HS C3N7) to 16 bp (HS C3N4, G18N1, and G18N2), with an average of 12 bp. The direct repeats were A-T rich with values ranging from 91% (10/11 bases) in HS C4N8 to 43% (3/7 bases) in HS C3N6 with an overall average of 67% A-T composition. In fact, the 5′ ends of the direct repeats were highly A-rich. Previous studies of the integration site specificity of Alu sequences (28) and other repetitive elements (29) have resulted in similar conclusions. Therefore the integration of HS subfamily members appears to occur in a manner typical of Alu sequences.

The oligo dA-tails of each subfamily member were also variable in length ranging from 11 bp (HS G18N1) to 37 bp (HS C4N2) with an average of 22 bp (Fig. 2). Although the length of the oligo-dA tails varied the composition did not. All of the HS subfamily members tails were comprised entirely of adenine residues except for HS C4N8 and P1N6. The tail of HS C4N8 contained one cytosine residue while the tail of HS P1N6 contained 3 guanine residues interspersed within it (Fig. 2). The simple nature of the sequence substitutions in these subfamily members suggests that Alu family members are initially inserted with a homogeneous oligo-dA tail. This initially perfect tail then apparently accumulates mutations and undergoes expansions and deletions at a high rate after retroposition and integration. This is supported by studies which showed that these types of mutations in the oligo-dA tail occur frequently as inherited polymorphisms within the human population (30).

Several of the HS subfamily members (HS P1N6, C3N7, C4N6, C4N8, G19N1 and C4N5) had 1–4 extra bases immediately adjacent to the 5′ end of the Alu sequence and 3′ of the direct repeat (Fig. 2). These bases may have resulted from reverse transcription of an Alu transcript which contains additional 5′ sequences (relative to the base 1 in the subfamily consensus), or from repair of the 5′ end of the insertion site. There is no consensus sequence for these additional bases therefore we favor the idea that they occurred during ligation/repair of the 5′-end of the Alu cDNA during integration. Two other HS subfamily members (HS C3N6 and C3N7) had additional bases between the 3′ end of their oligo-dA tails and direct repeat. Given the complex nature and length of these changes it is very unlikely that they arose as a result of point mutations in the oligo-dA tail. These changes may have resulted from an insertion at the 3′ end of the oligo-dA tail, or from the deletion of a portion of the 5′ direct repeat during integration. Further studies involving sequence analysis of the pre-integration sites will be required to determine the origin of both the 5′ and 3′ sequences.

A total of ten small internal insertion and deletion mutations were observed within the HS subfamily members reported here. Only four HS subfamily members (HS C3N7, MLVI-2, TPA 25 and HS C4N5) contained internal nucleotide insertions or deletions (Fig. 1). Two small 1−2 bp deletions as well as several point mutations are located within the right half (bases 213−277 relative the consensus in Fig. 1) of HS C3N7. This subfamily member also has an imperfect joint between the 3′ end of the oligo-dA tail and 3′ direct repeat (above). The small insertions, deletions and single nucleotide substitutions found within HS C3N7 may have resulted from mutations which occurred after integration of this subfamily member. Previous studies have shown that these types of mutational events (small insertions or deletions as well as single base substitutions) are characteristic of a number of environmental mutagens including low level ionizing radiation (31−33). The small deletion found in TPA 25 and HS C4N5 has previously been suggested as part of three tightly linked mutations from the HS subfamily consensus, forming an even more recent sub-subfamily of Alu family members (HS-2) (18, 34). Our finding that only two out of 20 HS subfamily members have the HS-2 changes suggests that the HS-2 subfamily is quite small, with only about 50 subfamily members located within the human genome. Only 19% (4/21) of the HS subfamily members contained internal nucleotide insertions or deletions. Therefore these type of events are not very common among recently inserted Alu sequences.

Inspection of the HS subfamily member nucleotide sequences shows that the majority (17/21) of the HS subfamily members are complete Alu copies (Fig. 2). Only 4/21 (19%) of the HS subfamily members were truncated at their 5′ terminus. Each of the truncated subfamily members (HS C3N1, C3N6, G15N2 and H3N1) is perfectly abutted by a direct repeat at the 5′ terminus, suggesting that these truncations may have occurred as a result of incomplete formation of the Alu cDNA during reverse transcription, during integration, or as a result of reverse transcription of an incomplete RNA.

DISCUSSION

The HS Alu subfamily members represent the most recently inserted group of Alu family members found within the human genome. Thus, they have both the lowest degree of overall nucleotide divergence between members (2.3%), and to the subfamily consensus sequence (1.1%). This makes them the most representative group of Alu sequences compared to the 'master' gene from which they were derived. The divergence from the consensus suggests that the subfamily has an average age of 2.8 million years. Thus, if we assume a linear expansion rate of the HS subfamily, the original 'master' HS subfamily member would have been created within the last 6 million years (2×2.8 million years). Several of the HS subfamily members (HS C2N4, C3N1, C4N4, C4N5, C4N6, and C4N8) have previously been shown not to predate the human/great ape divergence, also suggesting a recent origin for HS subfamily members (18). The human/great ape divergence is thought to have occurred 4−6 million years ago (26). Therefore, we feel that the estimated age based on intervening nucleotide sequence substitution rates is fairly accurate. However, the possibility that a small number of HS subfamily members predated the human/great ape divergence still exists as suggested by the calculated age of HS P1N5 (11.3 million years old). The low degree of nucleotide divergence also suggests that the individual HS subfamily members arose from a common source 'master' gene. The identity and exact nature of this source gene remains unknown. However, the consensus sequence for the HS subfamily (HS-2) should be the most accurate representation of the source gene currently undergoing amplification within the human genome.

The 3′ oligo-dA tails of individual subfamily members varied in length, but were almost exclusively composed of adenine. This type of random length perfect adenine rich tail is characteristic of post-transcriptional polyadenylation of RNA sequences. Based on similar data, a post-transcriptional origin for the oligo-dA tail of Alu sequences has been proposed previously (34). However, Alu sequences do not contain any known internal signals for post-transcriptional polyadenylation. The original model proposed for retroposition of Alu sequences suggests that the source 'master' gene has it's own 3′ oligo-A rich region (35, 36). The length of oligo-dA tails found within individual HS subfamily members would vary as a result of random self-priming for reverse transcription. The largest A tail found among HS subfamily members was 37 bp suggesting that the individual subfamily members were derived from a 'master' gene with an oligo-A tail at least 40 bp in length (allowing a few additional bases for self priming). The finding that the tails are originally so homogeneous, and not commonly composed of simple sequence repeats as are a number of the older Alu sequences, suggests that the model of Moos and Gallwitz (37), in which such simple sequence repeats are added at the time of insertion, is incorrect.

In older Alu family members, over one-third of the family members have no recognizable flanking, direct repeats (29). This could have been because they had inserted in the genome at a blunt end integration site (rather than at staggered nicks) or through mutational loss of the direct repeats. All of the HS members had perfect direct repeats of average length 12 bp. This suggests that some of the older Alu sequences had lost their direct repeats through mutation. Therefore, we conclude that all Alu sequences insert at staggered nicks (direct repeats), and that with time, the direct repeats are lost through mutation.

The rate of single nucleotide substitutions within individual HS subfamily members varied, with transition mutations (72%) occurring far more frequently than transversions (28%). This type of decay, favoring transition mutations, is consistent with that reported in a previous study of 50 randomly chosen human Alu sequences (28). The mutations within individual subfamily members appear to have occurred in a random manner similar to that found in a previous study (28). Therefore, we conclude that the HS subfamily members appear to evolve in a manner typical for older Alu sequences after integration within the genome.

The rate of nucleotide substitutions within HS subfamily members involving CpG dinucleotides was 9.2-fold higher than at non-CpG positions. These positions decay unidirectionally to form TpG or CpA residues as a result of the spontaneous deamination of 5-methyl cytosine (38, 39). Previous reports involving Alu sequences (15) and other pseudogenes (27) have shown about a 10-fold higher rate of decay for CpG positions.

Therefore the CpG decay rate within HS subfamily members appears typical for Alu sequences as well as other vertebrate pseudogenes. The CpG dinucleotides almost certainly represent mutation hotspots because of methylation. Different Alu family members have different levels of CpG changes. Thus, the Alu family members are almost certainly methylated in the genome. The numbers of CpG changes are low enough in the HS family members so that it is not certain whether this is statistical fluctuation, or whether perhaps individual Alu family members are subject to differing germ line methylation environments after insertion. In contrast to the individual Alu copies, the Alu consensus sequence, and therefore the Alu 'master' gene sequence, is very rich in CpG residues which have not changed much over primate evolution (above and 16). This suggests that the 'master' gene(s) is hypomethylated in the tissues in which germ line retroposition must occur.

ACKNOWLEDGEMENTS

We would like to thank Drs. Gary R. Daniels and Valerie K. Slagel for critical review of this manuscript. We would also like to thank Ms. Joelle Finley for cell culture assistance, and Mr. James Carlton of the L.S.U. Medical Center core laboratories for oligonucleotide synthesis. P.E.R. and T.D.D. were supported by the L.S.U. medical student summer research program. This work was supported by USPHS grant number RO1 HG00340 to P.L.D.

REFERENCES

1. Schmid, C.W., and Shen, C.-K.J. (1986) In MacIntyre, R.J. (ed.) Molecular Evolutionary Genetics. Plenum Press, N.Y., N.Y. pp. 323–358.
2. Weiner, A.M., Deininger, P.L., and Efstrdiatis, A. (1986) *Ann. Rev. Biochem.*, 55, 631–661.
3. Deininger, P.L. (1989) In Howe, M. and Berg, D. (eds.) Mobile DNA. ASM press, Washington D.C., pp. 619–636.
4. Deininger, P.L., and Daniels, G.R. (1986) *Trends Genet.*, 2, 76–80.
5. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedman, T., and Schmid, C.W. (1981) *J. Mol. Biol.*, **151**, 157–169.
6. Flemington, E.H., Bradshaw, H., Traina-Dorge, V., Slagel,V., and Deininger, P. (1987) *Gene*, 52, 267–277.
7. Lee, M.G.-S., Loomis, L., and Cowan, N.J. (1984) *Nucleic Acids Res.*, **12**, 5823–5836.
8. Ullu, E., Murphy, S., and Melli, M. (1982) *Cell*, **29**, 195–202.
9. Rogers, J. (1983) *Nature*, **301**, 460.
10. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw Jr., H., and Deininger, P.L. (1987) *Mol. Biol. Evol.*, 4, 19–29.
11. Willard, C., Nguyen, H.T., and Schmid, C.W. (1987) *J. Mol. Evol.*, **26**, 180–186.
12. Jurka, J., and Smith, T. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4775–4778.
13. Britten, R.J., Baron, W.F., Stout, D.B., and Davidson, E.H. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4770–4774.
14. Quentin, Y. (1988) *J. Mol. Evol.*, **27**, 194–202.
15. Labuda, D., and Striker, G. (1989) *Nucleic Acids Res.*, **17**, 2477–2491.
16. Deininger, P.L., and Slagel, V.K. (1988) *Mol. Cell. Biol.*, **8**, 4566–4569.
17. Matera, A.G., Hellmann, U., and Schmid, C.W. (1990) *Mol. Cell. Biol.*, **10**, 5424–5432.
18. Batzer, M.A., and Deininger, P.L. (1991) *Genomics*, In Press.
19. Friezner Degen, S.J., Rajput, B., and Reich, E. (1986) *J. Biol. Chem.*, **261**, 6972–6985.
20. Economou-Pachnis, A., and Tsichlis, P.N. (1985) *Nucleic Acids Res.*, **13**, 8379–8387.
21. Elder, J.T., Pan, J., Duncan, C.H., and Weissman, S.M. (1981) *Nucleic Acids Res.*, **9**, 1171–1189.
22. Perez-Stable, C., Ayres, T.M., and Shen, C.-K.J. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 5291–5295.
23. Gundelfinger, E.D., Carlo, M.D., Zopf, D., and Melli, M. (1984) *EMBO J.*, **3**, 2325–2332.
24. Zwieb, C. (1985) *Nucleic Acids Res.*, **13**, 6105–6124.
25. Siegel, V., and Walter, P. (1986) *Nature*, **320**, 81–84.
26. Miyamoto, M.M., Slightom, J.L., and Goodman, M. (1987) *Science*, **238**, 369–373.
27. Bulmer, M. (1986) *Mol. Biol. Evol.*, **3**, 322–329.
28. Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., and Matsubara, K. (1987) *Gene*, **53**, 1–10.
29. Daniels, G.R., and Deininger, P.L. (1985) *Nucleic Acids Res.*, **13**, 8939–8954.
30. Economou, E.P., Bergen, A.W., Warren, A.C., and Antonarkis, S.E. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 2951–2954.
31. Batzer, M.A., Tedeschi, B., Fossett, N.G., Tucker, A., Kilroy, G., Arbour, P., and Lee, W.R. (1988) *Mutation Res.*, **199**, 255–268.
32. Grimwade, B.G., Muskavitch, M.A.T., Welshens, W.J., Yedvobnik, B., and Artavanisj-Tsakovas, S. (1985) *Develop. Biol.*, **107**, 503–519.
33. Pastink, A., Schalet, A.P., Vreeken, C., Paradi, E., and Ecken, J.C.J. (1987) *Mutation Res.*, **177**, 101–115.
34. Matera, A.G., Hellmann, U., Hintz, M.F., and Schmid, C.W. (1990) *Nucleic Acids Res.*, In Press.
35. Jagadeeswaran, P., Forget, B.G., and Weissman, S.M. (1981) *Cell*, **26**, 141–142.
36. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T., and Gesteland, R.F. (1981) *Cell*, **26**, 11–17.
37. Moos, M., and Gallwitz, D. (1983) *E.M.B.O.J.*, **2**, 757–761.
38. Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) *Nature*, **244**, 775–780.
39. Bird, A.P. (1980) *Nucleic Acids Res.*, **8**, 1499–1504.